Capstone Project-4

Netflix Movies & TV Shows Clustering

Zeeshan Ahmad

Contents

- Problem Statement
- Data Summary
- Data Cleaning
- Exploratory Data Analysis
- Text Preprocessing
- Dimensionality Reduction using Principal Component Analysis
- Model Building
- Conclusion

Problem statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

In this project, we are required to do

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

Data Summary

The dataset has 7787 rows and 12 columns:

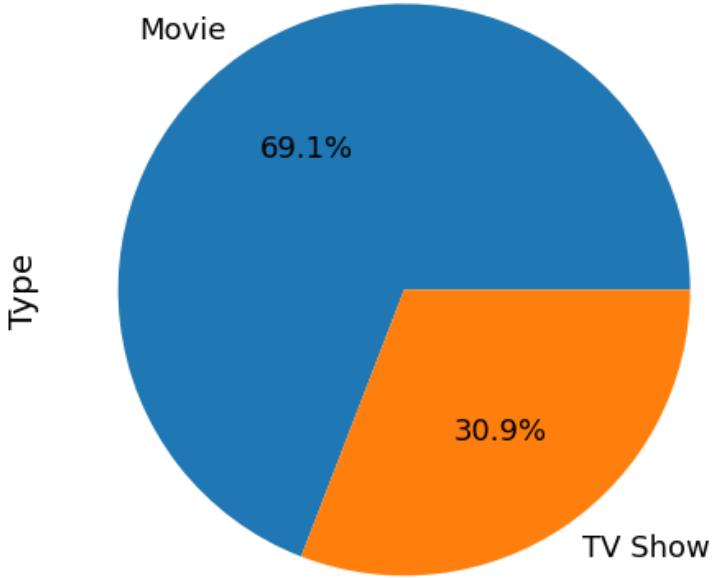
1. show_id : Unique ID for every Movie / Tv Show
2. type : Identifier - A Movie or TV Show
3. title : Title of the Movie / Tv Show
4. director : Director of the Movie
5. cast : Actors involved in the movie / show
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual Release year of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total Duration in minute or number of season
11. listed_in : Genres
12. description: The Summary description

Data Cleaning

- Filling the rows which has higher than 5% null and lower than 30% null values.
- Dropped the rows which has lower than 5% null values.
- Dropped the column which has null values higher than 30%
- Duplicate values: In our dataset we don’t have any duplicate records.

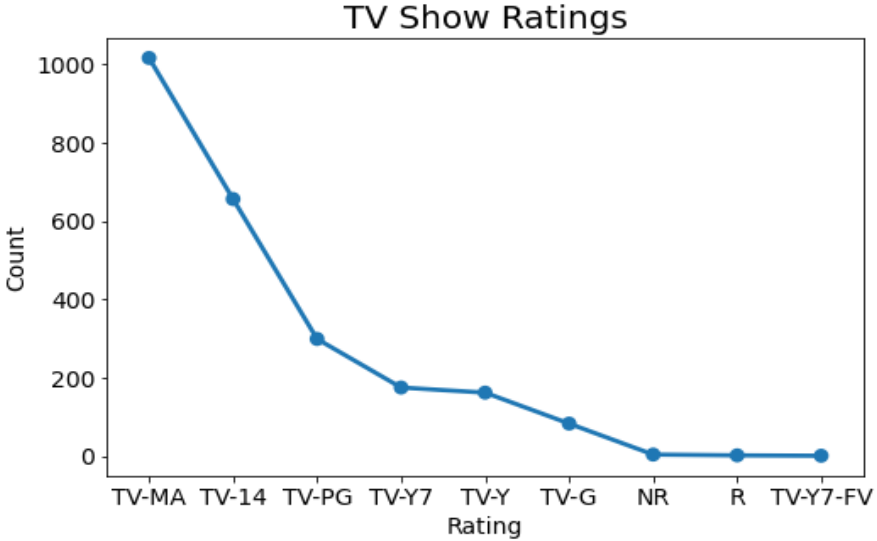
Exploratory Data Analysis

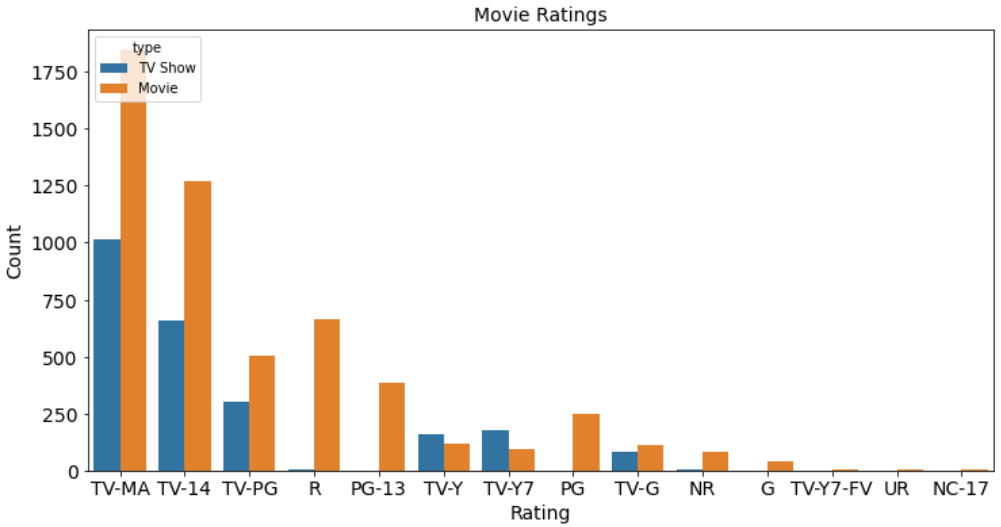
1. Type of Content



In dataset there is around 69% content as movies and remaining 31% as TV shows.

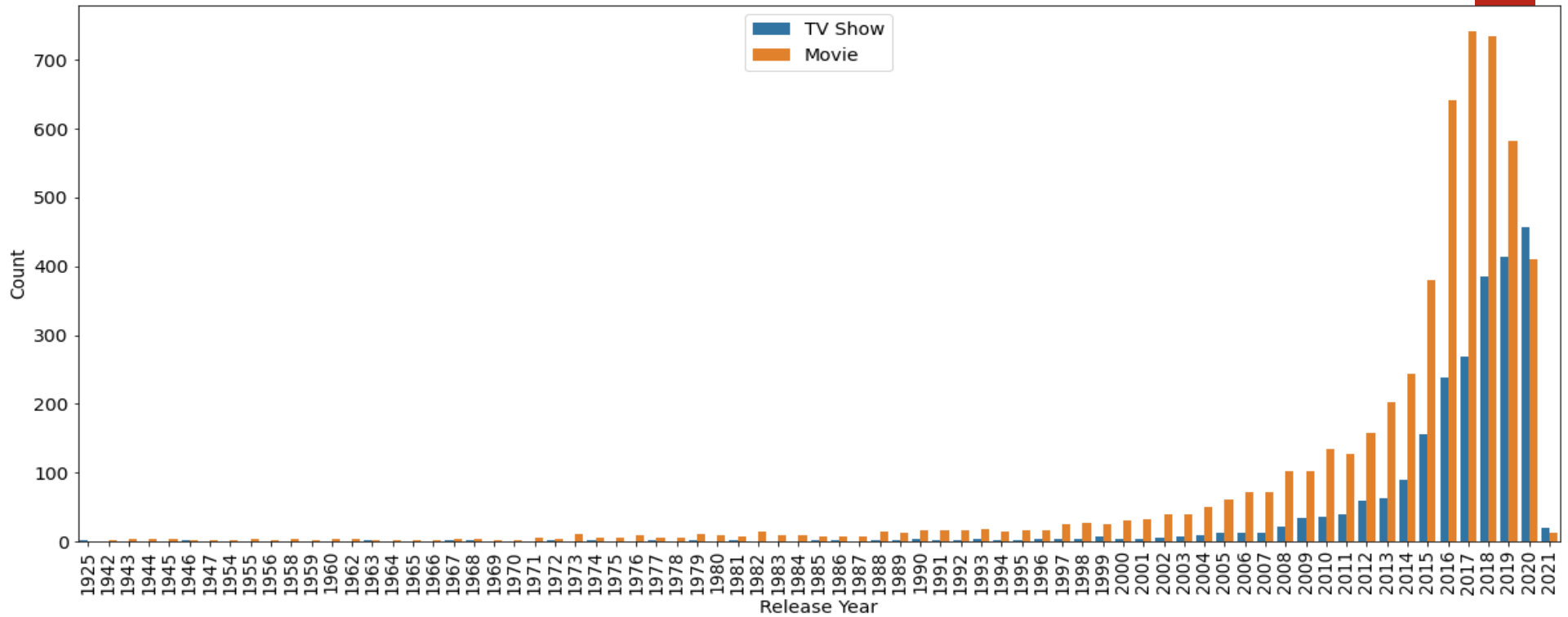
2. Rating of Content





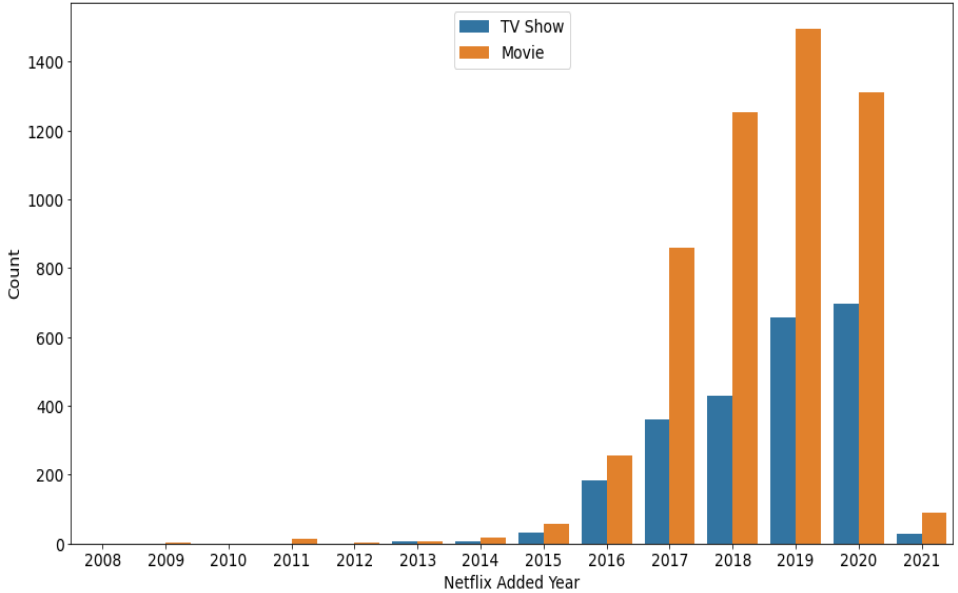
- Most of contents are in TV-MA, TV-14 and TV-PG ratings for both Movies and TV Shows.
- Contents for TV Shows are more than movies in rating TV-Y and TV-Y7
- The content for children and general audiences is less in Netflix eg. TV-Y7, TV-Y7-FV, G, etc

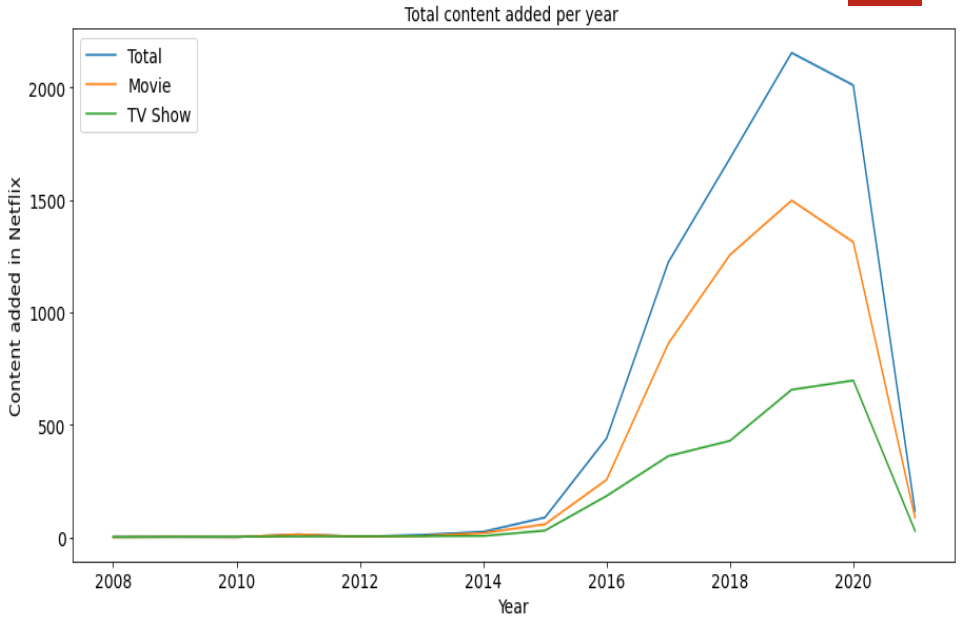
3. Release Year of Movies and TV Shows



- In year 2017 Count of Movies released are higher followed by 2018 and 2016.
- TV Shows are released higher in 2020 followed by 2019 and 2018.
- In year 2020 and 2021 TV shows are released more than movies
- After 2018 number of movies are decreased releasing.
- There is a significant drop in the number of movies and TV Shows produced after 2020.

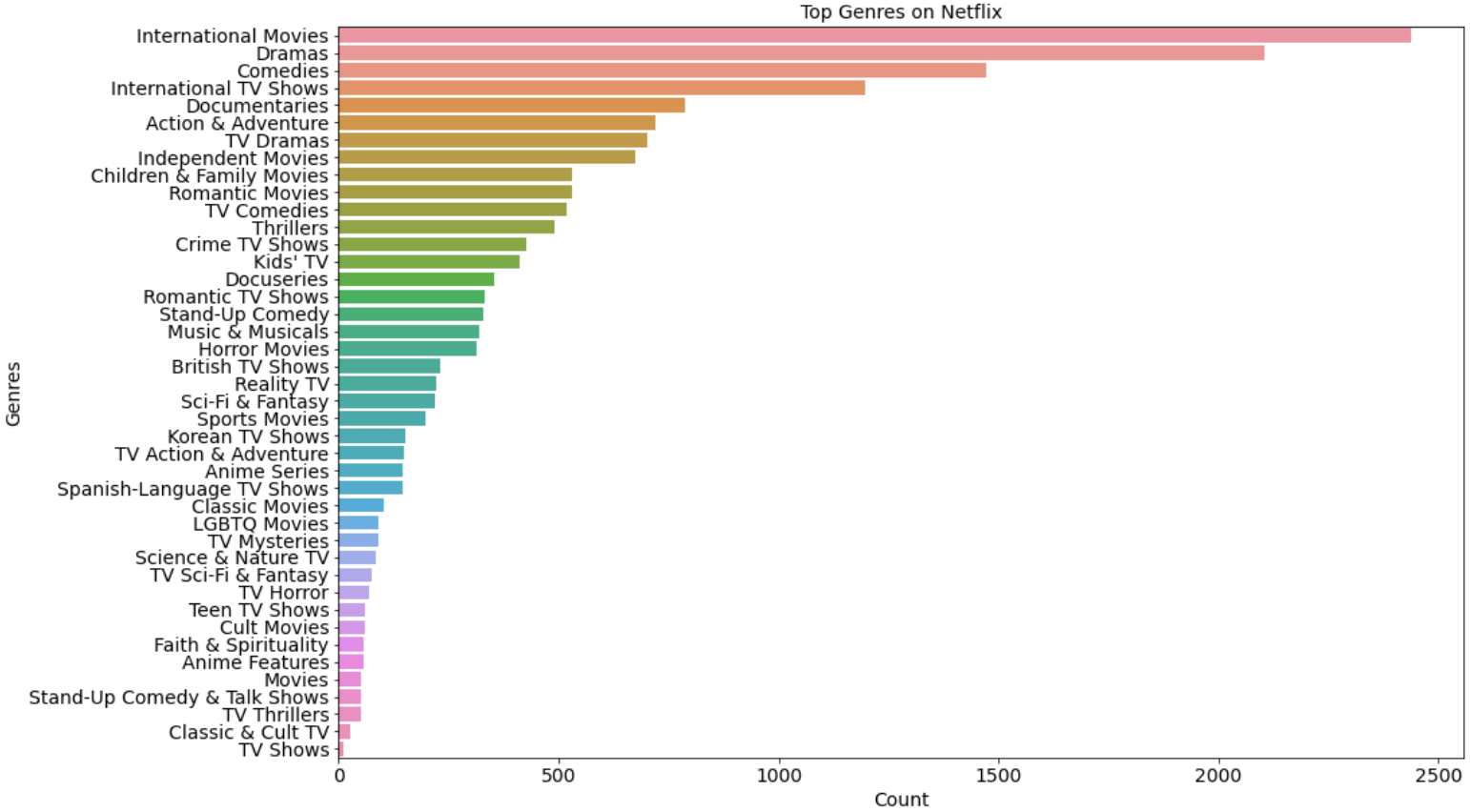
4. Content added year in Netflix





- The number of TV shows and movies added in 2019 and 2020 are maximum.
- Number of TV shows and movies decreased for year 2021.

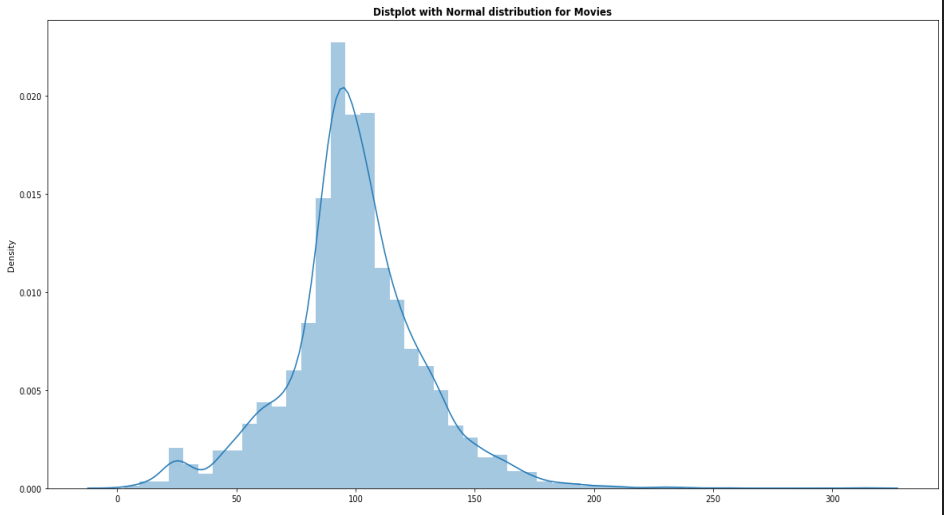
5. Top genres available in Netflix



- International movies, dramas, and comedies are the top three genres with the most content on Netflix.
- Rarest genres are available are Classic & Cult TV, TV shows

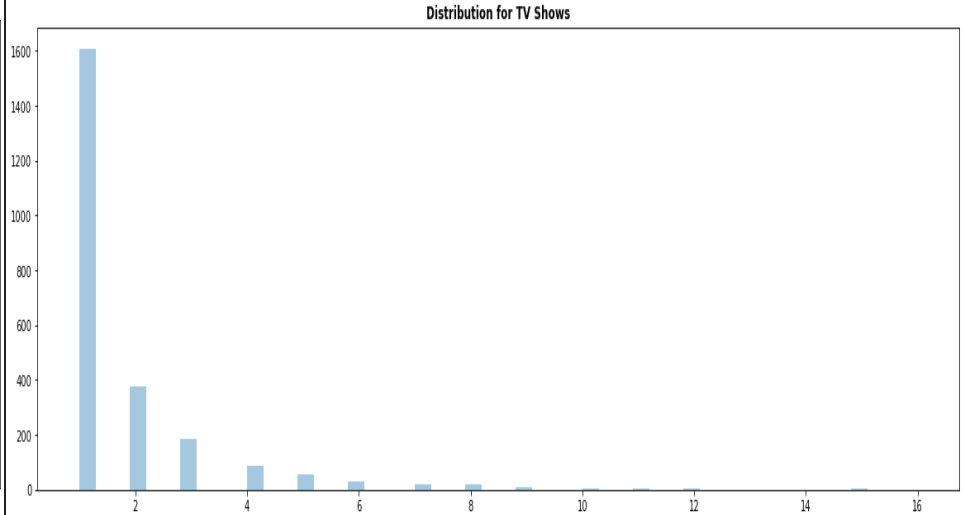
6. Duration

Distribution of movies duration



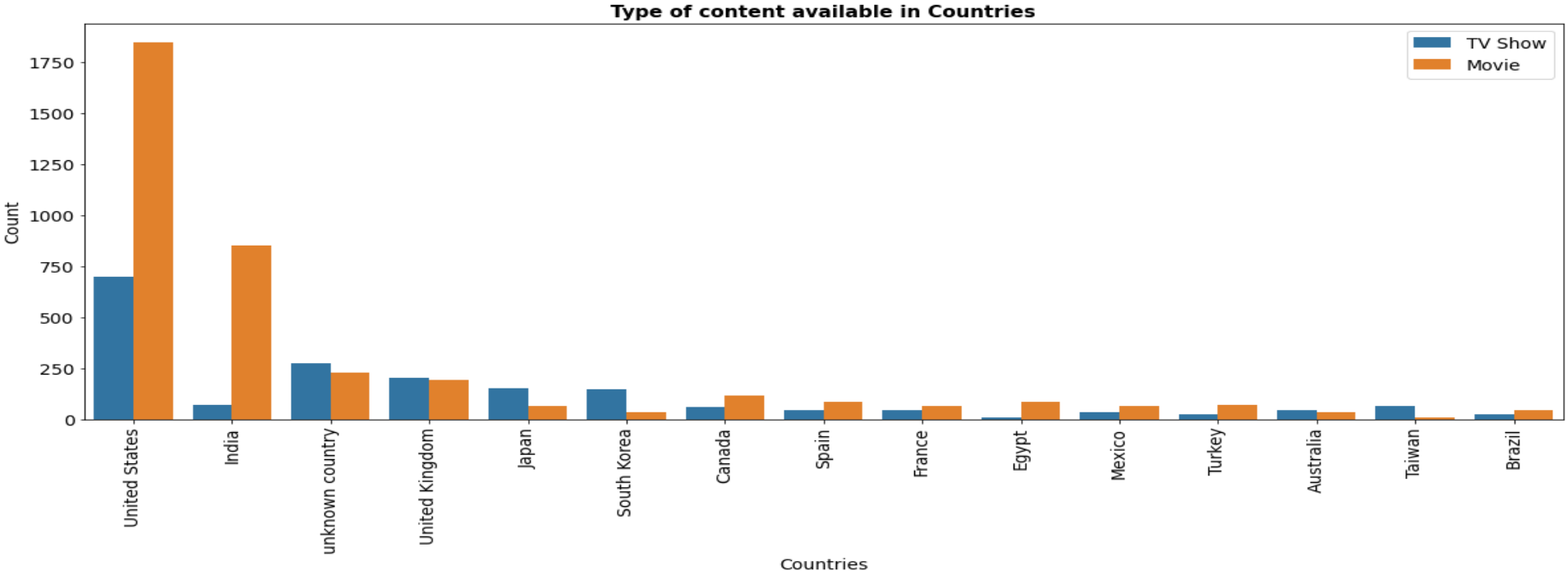
The majority of the films are between 80 and 120 minutes long.

Distribution of TV shows seasons



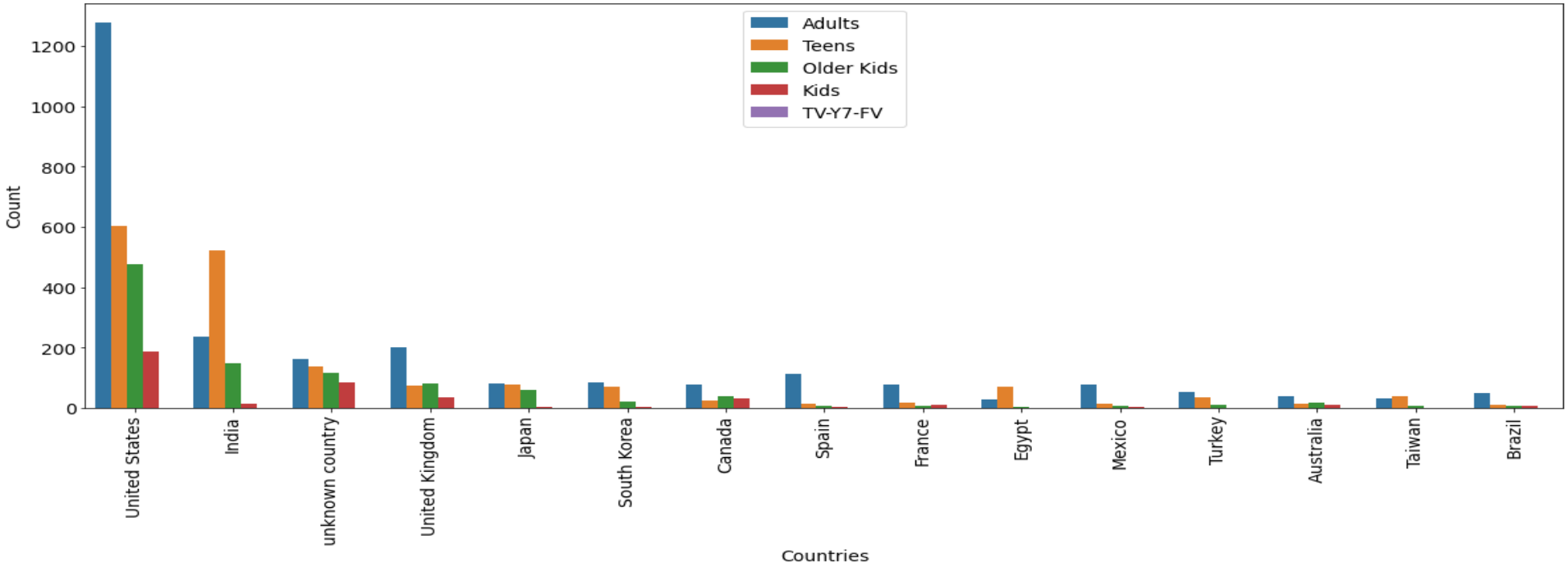
Highest number of tv_shows consisting of single season.

7. Type of content available in all countries



- United States have maximum movies in Netflix followed by India and United Kingdom.
- TV Shows are maximum in United States followed by United Kingdom and Japan.
- TV Shows in Egypt are less produced than all countries.
- Movies are less produced in Taiwan.

8. Target age of content available in all countries



- Content for adults are higher in United States followed by India, United Kingdom.
- Content for teens and kids are maximum in United States and India.
- Content is less in Brazil and Taiwan for every ages.

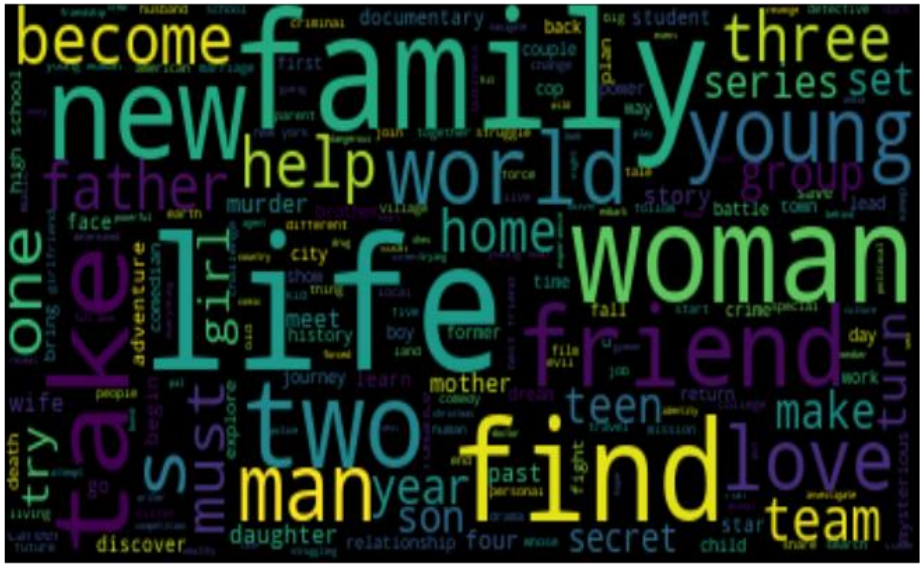
Text Pre-processing

1. Cleaning	2. Stop words	3. Tokenization	4.Lemmatization
• All words to lowercase • Removed Punctuation	• Removed Stopwords	• Split sentences into words	Grouped together the different inflected forms of a word so they can be analyzed as a single item

After all that transformed the text by using TFIDF Vectorizer

TF-IDF - Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector.

Word cloud



Description

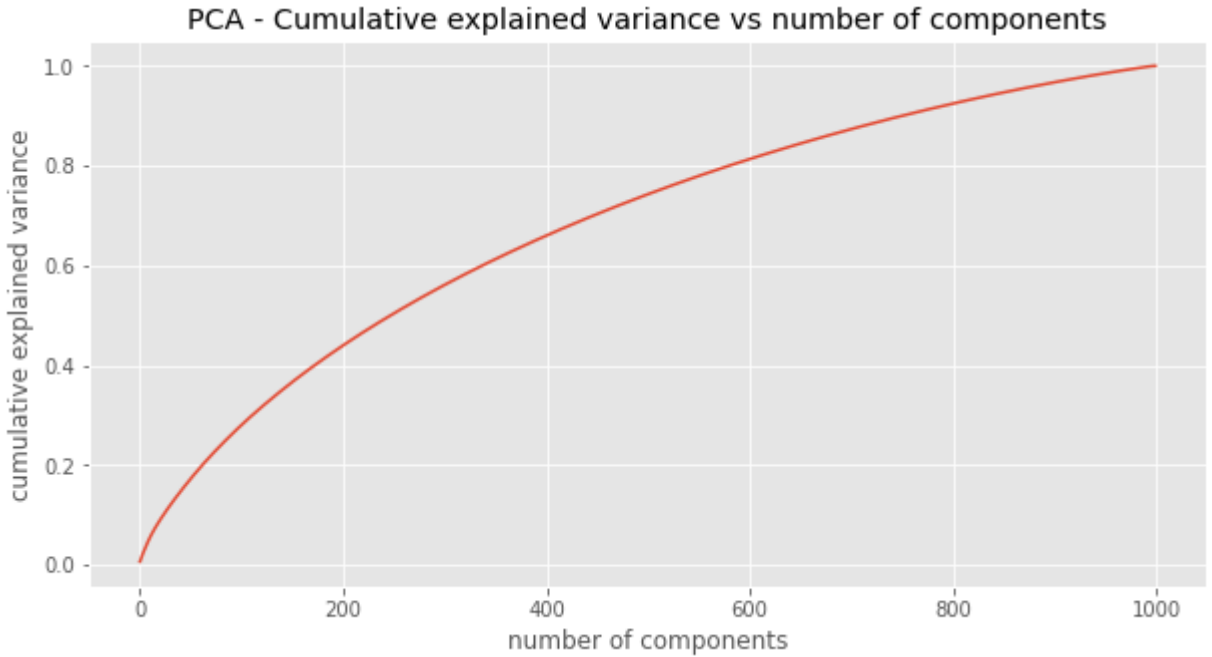
Most occurring words in the description of the tv shows and movies are Family, Friend, Love, Life, Woman, Man

Genres

Most occurring words in the genres are tv show, international movie, comedy, movie drama, comedy drama.

Dimensionality Reduction using Principal Component Analysis

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

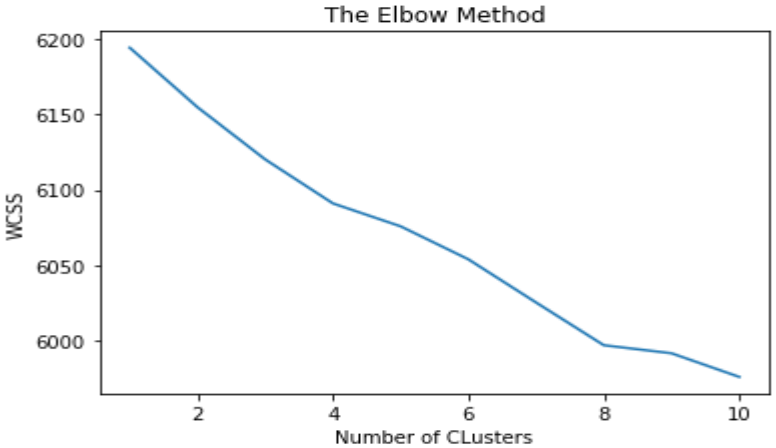


Here we can clearly spot that 80% variance is explained by 600 components only.

Model Building

1. K-Means Clustering

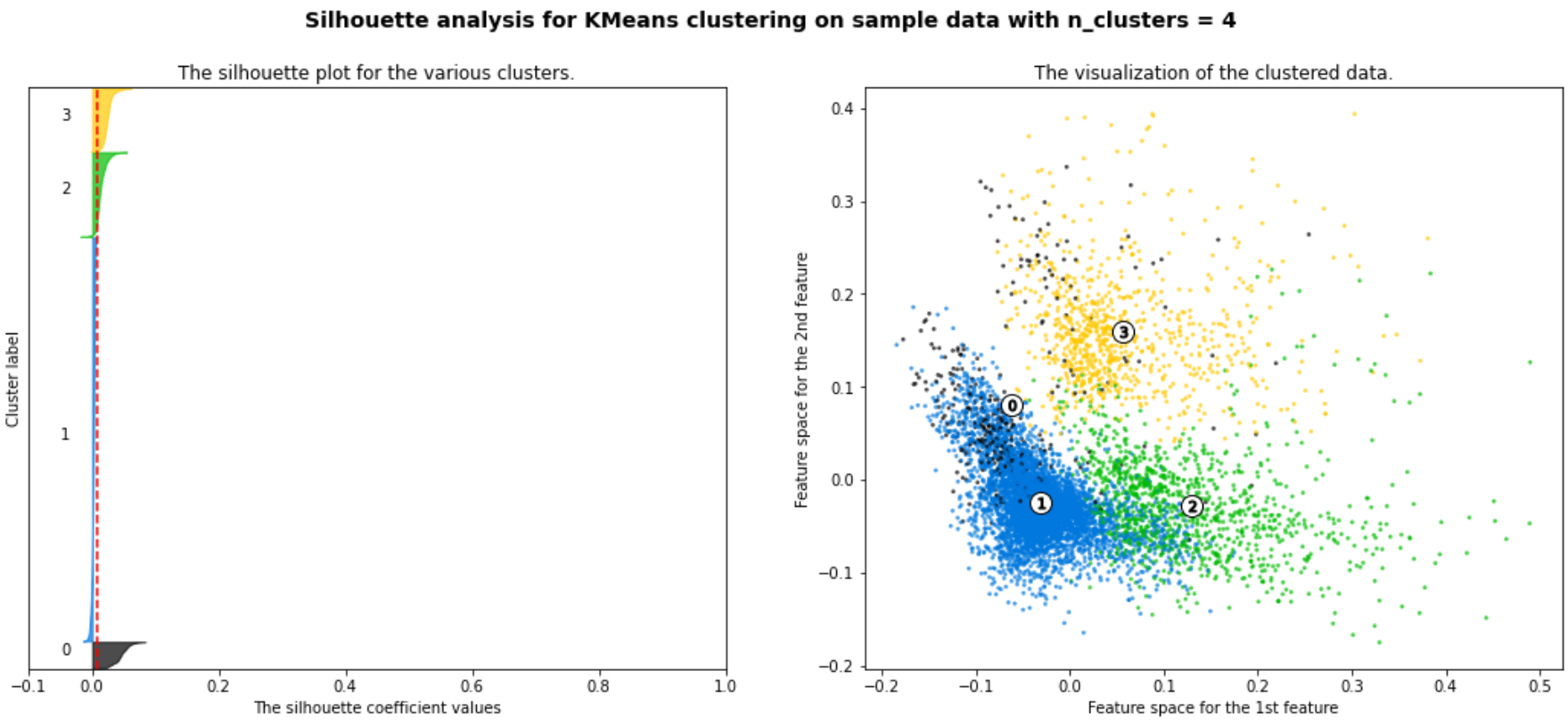
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similarproperties.




```
cluster: 2      Sillhoute: 0.0065
cluster: 3      Sillhoute: 0.0091
cluster: 4      Sillhoute: 0.0079
cluster: 5      Sillhoute: 0.0101
cluster: 6      Sillhoute: 0.0109
cluster: 7      Sillhoute: 0.0116
cluster: 8      Sillhoute: 0.0124
cluster: 9      Sillhoute: 0.0128
cluster: 10     Sillhoute: 0.0124
```

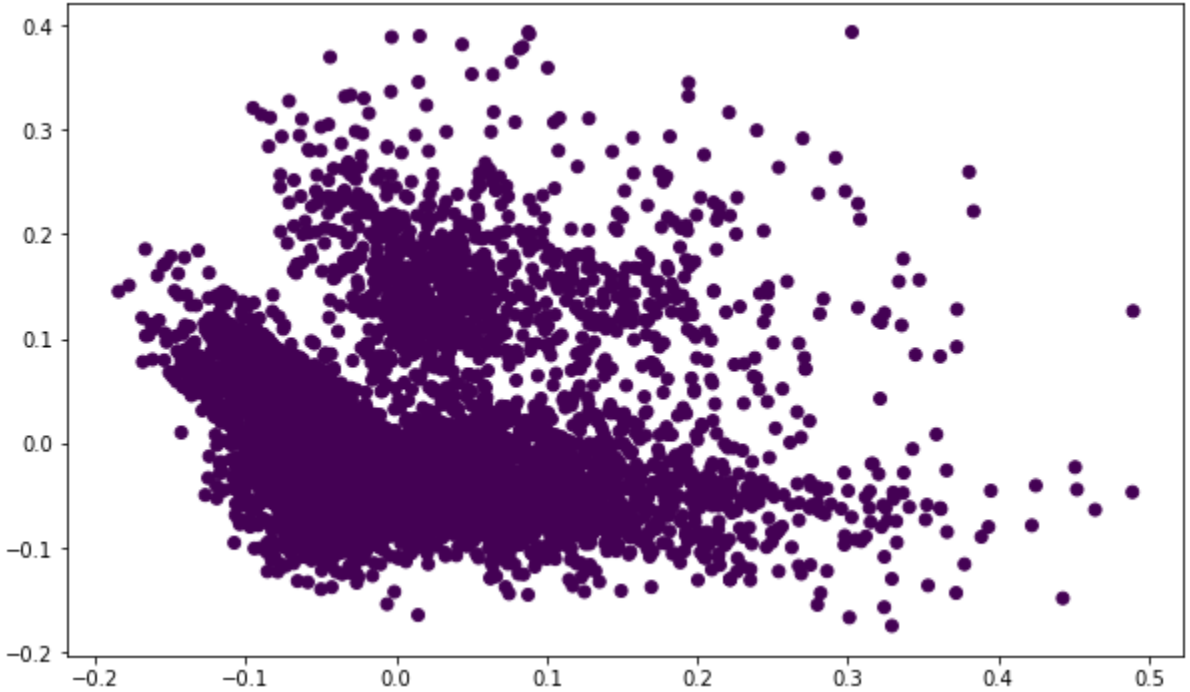
To find the number of clusters we used elbow method and Silhouette's score. After visualizing both, the best optimal number of clusters are 4.The elbow method uses sum of squared distances to choose an ideal value of no. of clusters.

Implementing K-Means Clustering Method



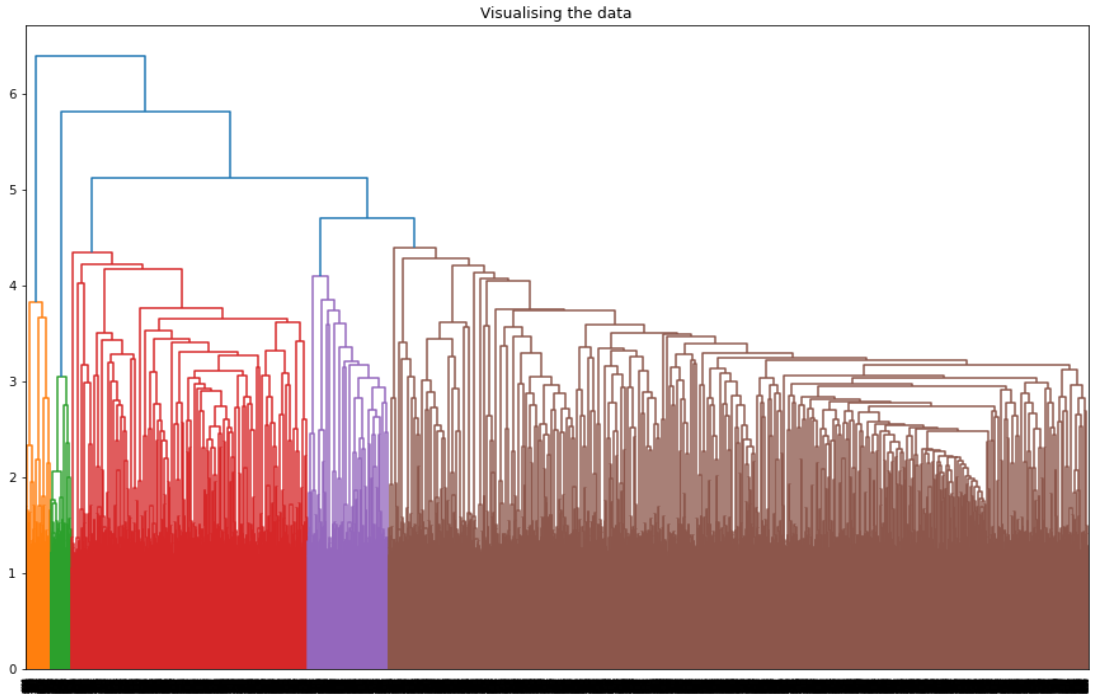
2. DBSCAN Clustering

DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. Given that DBSCAN is a density based clustering algorithm, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. DBSCAN can sort data into clusters of varying shapes as well, another strong advantage



3. Hierarchical Clustering

Finding number of clusters using Dendrogram after visualizing it Optimal Number of clusters I got was 4. Number of clusters will be number of vertical lines which are intersected by the line drawn using threshold.



A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

Conclusion

- In dataset there is around 69% content as movies and remaining 31% as TV shows.
- Most of contents are in TV-MA, TV-14 and TV-PG ratings for both Movies and TV Shows.
- Contents for TV Shows are more than movies in rating TV-Y and TV-Y7
- The content for children and general audiences is less in Netflix e.g. TV-Y7, TV-Y7-FV, G, etc.
- In year 2017 Count of Movies released are higher followed by 2018 and 2016.
- TV Shows are released higher in 2020 followed by 2019 and 2018.
- In year 2020 and 2021 TV shows are released more than movies
- There is a significant drop in the number of movies and TV Shows produced after 2020.
- The number of TV shows and movies added in 2019 and 2020 are maximum.
- International movies, dramas, and comedies are the top three genres with the most content on Netflix.
- United States have maximum movies in Netflix followed by India and United Kingdom.
- TV Shows are maximum in United States followed by United Kingdom and Japan.
- Content for adults are higher in United States followed by India, United Kingdom.
- Content for teens and kids are maximum in United States and India.
- Principal component analysis was performed in order to reduce the higher dimensionality
- Applied different clustering models K-means, hierarchical, DB Scan clustering on data we didn't get the best cluster arrangements.
- By applying different clustering algorithms to our dataset we get the optimal number of cluster is equal to 4

THANK YOU